Causal learning of biomolecular networks

Karen Sachs
Stanford University
Cells respond to their environment

Inside each cell is a molecular network

Secrete cytokines  Cell death  Proliferation

DNA damage  p53  E2F  p70 S6K  Akt  Bad  Bcl2  PI3K  MEK  Etk  cMyc  Stat  SAPK/JNK  ATF-2  C-JUN  CREB  Rb  p38 MAPK  JNK  ERK  MAPK  MAPK  MEK  MEKK  IKK  NF-kB  IκB-α
..which breaks down in disease states
Motive: Characterize normal, disease, drug.
Causal learning in signaling

Where does data come from?

- Technology
Causal learning in signaling

Where does data come from?

What causal connections appear?

• What happens?
• What can we see?
Causal learning in signaling

Where does data come from?

What causal connections appear?

What is needed for causal learning?

• Outstanding challenges
Causal learning in signaling

1. Where does data come from?
Samples are blended routinely

Tumor sample → Lab Blender! → Biological measurements
**Why single cell? (Biology perspective)**

*Innate* and *adaptive* branches of the Immune *system* communicate with each other to mount an effective immune response.

Cancer is a complex *system* with defined interdependent compartments.
Why single cell? (Stats perspective)

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>...</th>
<th>Xp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell1</td>
<td>120</td>
<td>...</td>
<td></td>
<td></td>
<td>77</td>
</tr>
<tr>
<td>.</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>.</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>.</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>CellN</td>
<td>12</td>
<td>...</td>
<td></td>
<td></td>
<td>34</td>
</tr>
</tbody>
</table>

High throughput data

Picture: John Albeck
From Phospho-molecular profiling to Signaling pathways
T-Lymphocyte Data

**Conditions** (multi-well format)
- Perturbation a
- Perturbation b
- Perturbation n

**11 Color Flow Cytometry**

- **Datasets of cells**
  - Condition 'a'
  - Condition 'b'
  - Condition...'n'

**Primary human T-Cells**
- 9 conditions
  - (6 Specific interventions)

**9 phosphoproteins, 2 phospholipids**
- 600 cells per condition
  - 5400 data-points
Accurate Network Inference

Model prediction validated with siRNA

[Sachs et al, Science 2005]
Flow Cytometry: Single Cell Analysis

Bendal et al, Science 2011
The Fluorescence Spectrum is Crowded

Fluorescent cytometry

- 6-8 parameters is “routine”
- 17 parameters has been reported
- Autofluorescence
- High background

Mass cytometry

- 100 discrete mass channels
- 38 parameters easily (58 soon)
- No compensation required
- Zero background
Antibody labels: isotopes of elements

Still dim limited! Currently working on this problem (talk offline)
How Do You Get 100 Channels From 35 Elements?

Each vertical bar is a different isotope that can be measured.
45-dimensional Single Cell Mass Cytometry

- Perturbations
- Integrate Pulse Scans Into 40+Dimensional Cell Events
- Stored In FCS Format

Isotopically enriched lanthanide ions (+3) x 6 polymers = 180 atoms per antibody

30-site chelating polymer

76,800 Mass Spectrum Scans Per Second
Cross-link Proteins

Permeabilize Cell Membrane

Nebulize To Single Cell Droplets

Ionize In Plasma (7500K)

ToF Mass Spec

Isotopically enriched lanthanide ions (+3)

30-site chelating polymer

x 6 polymers = 180 atoms per antibody

Tags
Workflow

Perturbations

Cross-link Proteins

Permeabilize Cell Membrane

Metal-chelated Antibody Stain

ToF Mass Spec

Ionize In Plasma (7500K)

Nebulize To Single Cell Droplets
Workflow: Measuring signaling by mass cytometry

Stimulate cells *in vitro*

Crosslink proteins

Permeabilize cell membrane

Stain with isotope tagged Abs

Nebulize single-cell droplets

Ionize (7500K)
Workflow: Measuring signaling by mass cytometry

1. Stimulate cells *in vitro*
2. Crosslink proteins
3. Permeabilize cell membrane
4. Stain with isotope tagged Abs
5. Nebulize single-cell droplets
6. Ionize (7500K)
7. Measure by TOF
Workflow: Measuring signaling by mass cytometry

1. Stimulate cells in vitro
2. Crosslink proteins
3. Permeabilize cell membrane
4. Stain with isotope tagged Abs
5. Nebulize single-cell droplets
6. Ionize (7500K)
7. Measure by TOF

Example graphs:
- Cell 1: Mass vs. CD45RA
- Cell 2: Mass vs. CD8
- pSTAT5
- pSTAT3
CyTOF: A prototype schematic

TOF (Time of Flight)

Nebulizer – Single cell droplets

ICP

Occurs at a rate of ~1000 cells per second

Fresh PBMc stained with 27 markers (mix I):

**Lymp CD4+T**
- CD45
- CD2
- CD3
- CD4

**Lymp B**
- CD45RA
- CD20
- CD45
- CD38
- CD71
- CD19
- CD40
Variation Across Calibrations and Instruments

Instrument 1

<table>
<thead>
<tr>
<th>Day A</th>
<th>Day B</th>
</tr>
</thead>
</table>

Instrument 2

<table>
<thead>
<tr>
<th>Day A</th>
<th>Day B</th>
</tr>
</thead>
</table>

Smoothed Bead Intensity

- Instrument 1:
  - Day A: [Graph]
  - Day B: [Graph]

- Instrument 2:
  - Day A: [Graph]
  - Day B: [Graph]
Bead normalization tames variation

Instrument 1

Day A

Instrument 2

Day A

Smoothed Bead Intensity

Instrument 1

Day B

Instrument 2

Day B

Smoothed Bead Intensity

Day A

Day B

time
Causal learning in signaling

1. Where does data come from?

2. What causal connections appear?
   - What happens?
   - What can we see?
Signaling 101
Signaling 101: Measure activated species

We measure these!

Cell response
Causal learning in signaling

Where does data come from?

What causal connections appear?

What is needed for causal learning?

• Outstanding challenges
Accurate Network Inference

Model prediction validated with siRNA

[Sachs et al, Science 2005]
Wait! What about..

- CPDs?

- Hidden variables?

- Cycles?

- Dynamics?


Sachs, Interface Focus 2013
Remaining challenges

- **CPDs?**

- **Cycles?**

- **Hidden variables?**

- **Dynamics?**
  - Sachs, Interface Focus 2013
Edges can be CPD dependent
CPDs need expressive power

- Multinomial?
- Linear?
- Other?
- GP (J. Mooij)
Remaining challenges

- **CPDs?**

- **Hidden variables?**

- **Cycles?**

- **Dynamics?**
  - Sachs, Interface Focus 2013
Comparative signaling
Elucidated edge is supported in the data
Workflow for adding hidden variables

1. **Workflow**
   - pSTAT1
   - CREB
   - **Statistical inference**
   - **In silico predictions**
     - Literature & curated DBs
     - **Scansite**
     - **In silico predictions**
   - **Putative directed paths and common ancestors**
   - Rank and select most probable path(s)

2. **Putative directed paths and common ancestors**
   - A → B → C
   - A → D → F → C
   - A ↔ G → C

3. **Networks**
   - IGF1R(3480) → PHOS(Y1229) → MUC1(1) → PHOS(SNR) → IKBA(4792)
   - IGF1R(3480) → PHOS(Y1229) → MUC1(1) → PHOS(SNR) → IKBA(4792)
   - CDK1(1) → PHOS(SIGNED_KINOME) → ABL1(25)
   - CREB
   - Stat1

4. **Statistical inference**
   - pSTAT1
   - CREB
   - **Statistical inference**
Uncovering underlying path

Branch active in Naïve
Not active in Memory
Remaining challenges

- CPDs?
- Cycles?
- Hidden variables?
- Dynamics?

Special case


Sachs, Interface Focus 2013
Remaining challenges

- CPDs?
- Cycles?
- Hidden variables?
- Dynamics?


Sachs, Interface Focus 2013
Cyclic structure learning

Itani and Sachs et al, JMLR Proc 2008
ODE model for realistic synthetic data

Jonathan Fitzgerald, Birgit Schoeberl
Merrimack Pharmaceuticals
Extra edge?

Ground truth

Model results
Zoom in on signaling

IGF

IGFR

IRS → PI3K → AKT → mTOR

We measure these!

Ground truth
What happens when [A] is limiting?

Note: Not revealed by activity inhibition!

Zoom in on signaling
Extra edge: causal via competition

Ground truth

Model results
Remaining challenges

- **CPDs?**

  \[ \text{A} \rightarrow \text{B} \]

- **Cycles?**


  \[ \text{GRB2/SOS} \rightarrow \text{Ras} \rightarrow \text{Raf} \rightarrow \text{MEK} \rightarrow \text{Erk} \]

- **Hidden variables?**

  \[ \text{A} \rightarrow \text{B} \rightarrow \text{D} \rightarrow \text{A} \rightarrow \text{C} \rightarrow \text{D} \]

- **Dynamics?**

  Sachs, Interface Focus 2013
Cyclic structure learning

Itani and Sachs et al, JMLR Proc 2008
Dynamics can confound causality:

**Example**

\[ A(t) \] randomly reset to 0 or 1

\[ B = A(t-1), \]

\[ C(t) = A(t) \lor A(t-1) \]
Bio Example

\[ A \xrightarrow{t} B \xrightarrow{t+\varepsilon} C \]

\[ A \xrightarrow{} B \rightarrow C \]
Bio Example: C depends on the *history* of A
We are generally not in SS

Schoeberl et al 2002
Avoid dependence on A’s history? (How?)
Hold A constant!
Algorithm for reducing noncausal edges

• Avoid dependence on history by learning from multi-inhibited conditions
• Some formalized results (see past talk)
• Continuing work based on feedback (work in progress)
Reconstruction in T Cells

Standard BN

Combined inhibitions
Other approaches?
Other approaches?
Other approaches?

How much fits into one snapshot?
Bonuss challenge

- CPDs?
- Cycles?
- Hidden variables?

Variable noise

Sachs, Interface Focus 2013
Bonus challenge: Variable noise

B has higher measurement noise

$A \rightarrow B \rightarrow C$

Ground truth
Bonus challenge: Variable noise

B has higher measurement noise

Ground truth
Bonus challenge: Variable noise

B has higher measurement noise

A → B → C

Ground truth

A → C → B

Learned model
Acknowledgements

Team Stanford
Garry Nolan
Tiffany Chen
Tim Lee
Peng Qiu
Sean Bendall
Erin Simonds

Team Berkeley, UMN, Merrimack, Harvard
Wendy Fantl
Tyler Burns
Astraea Jager
Andrew Gentles
Sylvia Plevritis
Mingyu Chung
Mohammed AlQuraishi
Birgit Schoeberl
Jonathan Fitzgerald
Solomon Itani
Claire Tomlin
Zohar Sachs

Thank you!
karensachs@stanford.edu

Funding
LLS post doctoral fellowship

The Leukemia & Lymphoma Society®
Fighting Blood Cancers